Statistica della Formazione

Slides 4

A.A. 2020-2021

Docente: ANNA LINA SARRA

Modulo 1: elementi di statistica descrittiva

- **Le misure di variabilità**

Variabilità

Per variabilità si intende l'attitudine dei fenomeni, naturali e sociali, a manifestarsi in modi differenti.

- **La variabilità è l'attitudine di un carattere a presentare modalità differenti nel collettivo in esame.**
- **La distribuzione di un carattere presenta variabilità nulla se su tutte le unità statistiche si rileva la stessa modalità. In tal caso tutti gli indici di variabilità assumono valore zero.**

Variabilità per caratteri di natura qualsiasi
Eterogeneità

- Misura la variabilità nelle distribuzioni secondo **caratteri qualitativi** o quantitativi

  - **minima eterogeneità (massima omogeneità):** tutte le unità del collettivo hanno la stessa modalità del carattere

  - **massima eterogeneità** (**minima omogeneità)** le modalità presentano tutte la stessa frequenza.

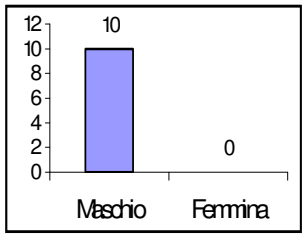
Indice di eterogeneità $E = 1 - \sum_{i=1}^{k} f_i^2$

valore minimo 0 (*eterogeneità nulla*)

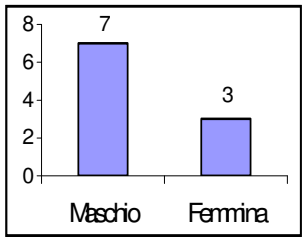
valore massimo $\left(\frac{k-1}{k}\right)$ (*eterogeneità massima*)

Eterogeneità: esempi

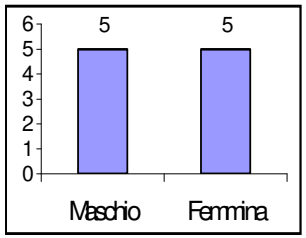
x_i	n_i
Maschio	10
Femmina	0
Totale	10



x_i	n_i
Maschio	7
Femmina	3
Totale	10



x_i	n_i
Maschio	5
Femmina	5
Totale	10



Eterogeneità nulla

Eterogeneità massima

x_i	n_i	f_i	f_i^2
Maschio	10	1	1
Femmina	0	0	0
Totale	**10**	**1**	**1**

$$E = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 1 = 0$$

x_i	n_i	f_i	f_i^2
Maschio	7	0,7	0,49
Femmina	3	0,3	0,09
Totale	**10**	**1**	**0,58**

$$E = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0.58 = 0.42$$

x_i	n_i	f_i	f_i^2
Maschio	5	0,5	0,25
Femmina	5	0,5	0,25
Totale	**10**	**1**	**0,5**

$$E = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0.5 = 0.5$$

Indice di eterogeneità: esempio

$$E = 1 - \sum_{i=1}^{k} f_i^2$$

Tipo di delitto	n_i	f_i	f_i^2
furti con strappo	1808	0.028	0.0008
furti con destrezza	7513	0.118	0.0138
furti in abitazioni	14756	0.231	0.0534
furti in esercizi commerciali	35626	0.558	0.3112
furti in auto in sosta	4163	0.065	0.0042
Totale	63866	1	0.3834

$$E = 1 - 0.3834 = 0.6166$$

$$\max(E) = \frac{k-1}{k} = \frac{5-1}{5} = 0.80$$

$$e = \frac{E}{\max E} = \frac{0.6166}{0.80} = 0.77$$

L'indice di eterogeneità relativo si calcola dividendo l'indice di eterogeneità per il suo massimo

Variabilità per caratteri quantitativi: Dispersione rispetto alla media aritmetica

- La dispersione rappresenta la variabilità delle modalità presentata del carattere rispetto ad un valore di sintesi scelto, ad esempio la media aritmetica della distribuzione
- Gli indici che danno una misura della variabilità dei valori della distribuzione rispetto alla media aritmetica sono lo **scostamento semplice medio** e lo **scostamento quadratico medio**

Misura della variabilità: scostamento semplice medio

- Lo **scostamento semplice medio è** la **media aritmetica degli scarti dalla media presi in valore assoluto**

Distribuzione disaggregata

$$S_\mu = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots + |x_N - \mu|}{N}$$

$$= \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|.$$

Distribuzione di frequenza

$$S_\mu = \frac{|x_1 - \mu| \cdot n_1 + |x_2 - \mu| \cdot n_2 + \cdots + |x_k - \mu| \cdot n_k}{N}$$

$$= \frac{1}{N} \sum_{i=1}^{k} |x_i - \mu| \cdot n_i = \sum_{i=1}^{k} |x_i - \mu| \cdot f_i$$

Scostamento semplice medio per la distribuzione disaggregata: calcolo

$$S_{\mu} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \mu|$$

x_i	$\lvert x_i - \mu \rvert$
0	7.14
5	2.14
6	1.14
8	0.86
9	1.86
10	2.86
12	4.86
Totale	20.86

μ=7.14 *Provvedimenti disclipinari*

Numero di provvedimenti disciplinari registrati in un campione di 7 scuole

0, 5, 6, 8, 9, 10, 12

❑ *Scostamento semplice medio*:

$$S_{\mu} = \frac{|0-7.14|+|5-7.14|+\cdots+|12-7.14|}{7}$$

$$= \frac{20.86}{7} = 2.98$$ *Provvedimenti disciplinari*

Scostamento semplice medio per una distribuzione di frequenze : calcolo

$$S_\mu = \frac{1}{N}\sum_{i=1}^{k} |x_i - \mu| \cdot n_i$$

N. di provvedimenti disciplinari

x_i	n_i	$x_i * n_i$	$\lvert x_i - \mu \rvert$	$\lvert x_i - \mu \rvert \cdot n_i$
1	1	1	4.42	4.42
2	2	4	3.42	13.68
3	3	9	2.42	21.78
4	3	12	1.42	17.04
5	1	5	0.42	2.1
6	2	12	0.58	6.96
7	1	7	1.58	11.06
8	2	16	2.58	41.28
9	3	27	3.58	96.66
10	1	10	4.58	45.8
totale	19	103		260.78

- Media aritmetica

$$\mu = \frac{103}{19} = 5.42$$

- *Scostamento semplice medio*

$$S_\mu = \frac{|1-5.42|\cdot 1 + \cdots + |1-5.42|\cdot 10}{19}$$

$$= \frac{260.78}{19} = 13.73$$

Misura della variabilità: scostamento quadratico medio

Lo **scostamento quadratico medio** o **deviazione standard** è la **media quadratica degli scarti dalla media**

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

Distribuzione disaggregata

$$\sigma = \sqrt{\frac{(x_1-\mu)^2 n_1+(x_2-\mu)^2 n_2+\cdots+(x_k-\mu)^2 n_k}{N}} = \sqrt{\frac{\sum_{i=1}^{k}(x_i-\mu)^2 n_i}{N}}$$

Distribuzione di frequenza

Formula alternativa

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} x_i^2 n_i - \mu^2}$$

Scostamento quadratico medio per la distribuzione disaggregata: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

x_i	$(x_i - \mu)^2$
0	51.02
5	4.59
6	1.31
8	0.73
9	3.45
10	8.16
12	23.59
Totale	92.86

Numero di provvedimenti disciplinari registrati in un campione di 7 scuole

0, 5, 6, 8, 9, 10, 12

❑ *Scostamento quadratico medio:*

$$\sigma = \sqrt{\frac{(0-7.14)^2 + (5-7.14)^2 + \cdots + (12-7.14)^2}{7}}$$

$$= \sqrt{\frac{92.86}{7}} = 3.64$$

Provvedimenti disciplinari

$\mu = 7.14$ *Provvedimenti disciplinari*

σ per una distribuzione di frequenze a modalità singole: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \mu)^2 n_i}{N}}$$

N. di provvedimenti disciplinari

x_i	n_i	$x_i * n_i$	$(xi - \mu)^2$	$(x_i - \mu)^2 \cdot n_i$
1	1	1	19.54	19.54
2	2	4	11.70	46.79
3	3	9	5.86	52.71
4	3	12	2.02	24.20
5	1	5	0.18	0.88
6	2	12	0.34	4.04
7	1	7	2.50	17.47
8	2	16	6.66	106.50
9	3	27	12.82	346.04
10	1	10	20.98	209.76
totale	19	103		827.93

- Media aritmetica

$$\mu = \frac{103}{19} = 5.42$$

- *Scostamento quadratico medio*

$$\sigma = \sqrt{\frac{(1-5.42)^2 \cdot 1 + \cdots + (1-5.42)^2 \cdot 10}{19}} =$$

$$= \sqrt{\frac{827.93}{19}} = 6.6$$

Proprietà degli indici S_μ e σ

- Assumono il valore 0 nel caso di assenza di variabilità (*tutte le unità statistiche presentano la stessa modalità del carattere*)
- **Sono espressi nella stessa unità di misura del carattere considerato**
- Non cambiano se a ciascun termine della distribuzione si aggiunge una quantità costante positiva o negativa
- La moltiplicazione di ciascun termine della distribuzione per una costante, positiva o negativa, ha come conseguenza la moltiplicazione degli indici per il valore assoluto della costante

Varianza

Il quadrato della deviazione standard si chiama varianza

(è espressa in unità di misura del carattere al quadrato).

$$\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$ *Distribuzione disaggregata*

$$\sigma^2 = \frac{(x_1-\mu)^2 n_1+(x_2-\mu)^2 n_2+\cdots+(x_k-\mu)^2 n_k}{N} = \frac{\sum_{i=1}^{k}(x_i-\mu)^2 n_i}{N}$$ *Distribuzione di frequenza*

Formula alternativa

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}x_i^2 n_i-\mu^2$$

Devianza

La somma dei quadrati degli scarti dalla media (numeratore della varianza) si chiama devianza

(*è espressa in* ***unità di misura del carattere al quadrato***)

$$D = (x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2 = \sum_{i=1}^{N} (x_i - \mu)^2$$

Distribuzione disaggregata

$$D = (x_1 - \mu)^2 n_1 + (x_2 - \mu)^2 n_2 + \cdots + (x_k - \mu)^2 n_k = \sum_{i=1}^{k} (x_i - \mu)^2 n_i$$

Distribuzione di frequenza

Varianza e devianza per la distribuzione disaggregata: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

x_i	$(x_i - \mu)^2$
0	51.02
5	4.59
6	1.31
8	0.73
9	3.45
10	8.16
12	23.59
Totale	92.86

Numero di provvedimenti disciplinari in un campione di 7 scuole

0, 5, 6, 8, 9, 10, 12

$$\sigma = \sqrt{\frac{92.86}{7}} = 3.64$$ ***Provvedimenti disciplinari***

$$\sigma^2 = \frac{92.86}{7} = 13.27$$ ***Provvedimenti disciplinari ^{2}***

$$devianza = 92.86$$ ***Provvedimenti disciplinari ^{2}***

μ=7.14 *Provvedimenti disciplinari*

Indici di variabilità relativa

Il problema si pone nel confrontare gli indici di variabilità di 2 o più distribuzioni *diverse*

Indice di variabilità relativa: numero puro a-dimensionale che elimina l'influenza dell'unità di misura e dell'intensità media

Indici di variabilità percentuali

- Si chiama **indice di variabilità percentuale** il rapporto, moltiplicato per 100, tra un indice di variabilità assoluto e la media aritmetica.
- Particolare rilievo per le applicazioni ha **coefficiente di variazione**

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Coefficiente di variazione: esempio

È maggiormente variabile il reddito medio annuo familiare o il numero di componenti della famiglia?

Reddito *(in migliaia)*	Media : 33364 euro Deviazione standard : 24636 euro
Numero di componenti	Media : 2.77 componenti Deviazione standard : 1.31 componenti

$$\boldsymbol{CV(reddito) = \frac{24636\ euro}{33364\ euro}100 = 74\%}$$

$$\boldsymbol{CV(componenti) = \frac{1.31\ componenti}{2.77\ componenti}100 = 47\%}$$

Coefficiente di variazione: esempio

Il reddito medio annuo è maggiormente variabile nell'insieme delle famiglie con 2 o con 4 componenti?

Reddito *(2 componenti)*	Media : 24451 euro Deviazione standard : 21218 euro
Reddito *(4 componenti)*	Media : 49260 euro Deviazione standard : 26050 euro

$$\boldsymbol{CV(reddito-2componenti)=\frac{21218\,euro}{24451\,euro}100=86.4\%}$$

$$\boldsymbol{CV(reddito-4componenti)=\frac{26050\,euro}{49260\,euro}100=52.9\%}$$

Campo di variazione e differenza interquartile

Sia $x_1, x_2, \ldots, x_N$ una distribuzione disaggregata. Sia $y_1, y_2, \ldots, y_N$ la stessa distribuzione con i termini disposti in ordine non decrescente.

campo di variazione

$$\Delta_c = y_N - y_1$$

Differenza tra massimo e minimo

Intervallo interquartile

$$\Delta_q = q_3 - q_1$$

Differenza tra terzo e primo quartile

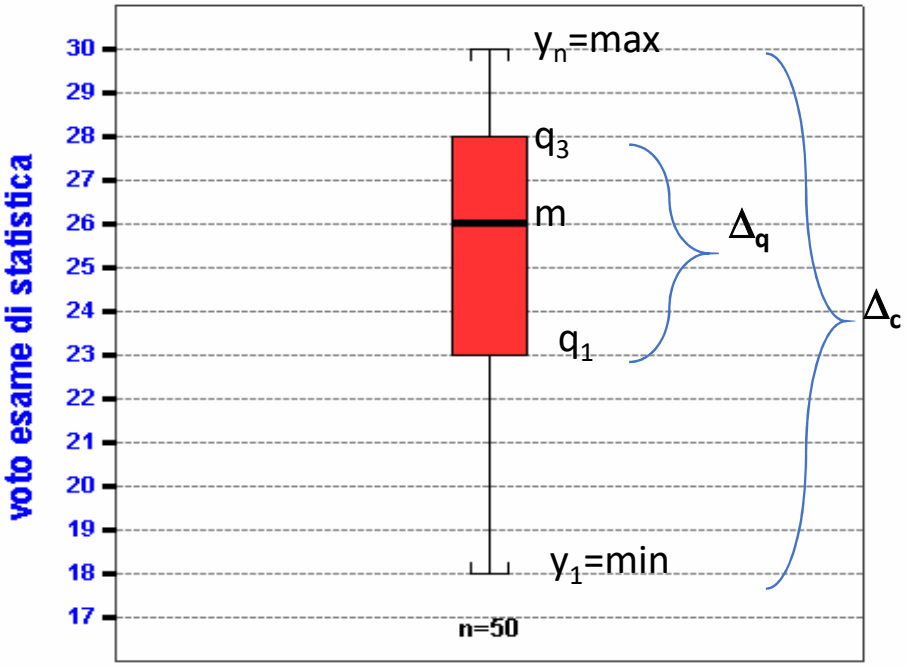
IL BOX-PLOT

- Una descrizione sintetica e abbastanza completa di una distribuzione di frequenze secondo un carattere quantitativo è data dal **box-plot**; questo è un riassunto a cinque numeri.
- I numeri sono i seguenti:
  - - la mediana (che dà informazioni sulla tendenza centrale)
  - - il primo e terzo quartile (la cui differenza dà informazioni sulla variabilità)
  - - i due estremi (la modalità più grande e la modalità più piccola)
- Questi numeri forniscono una descrizione sintetica di un insieme di dati anche quando il numero di unità osservate è elevato.

Box plot o Diagramma a scatola

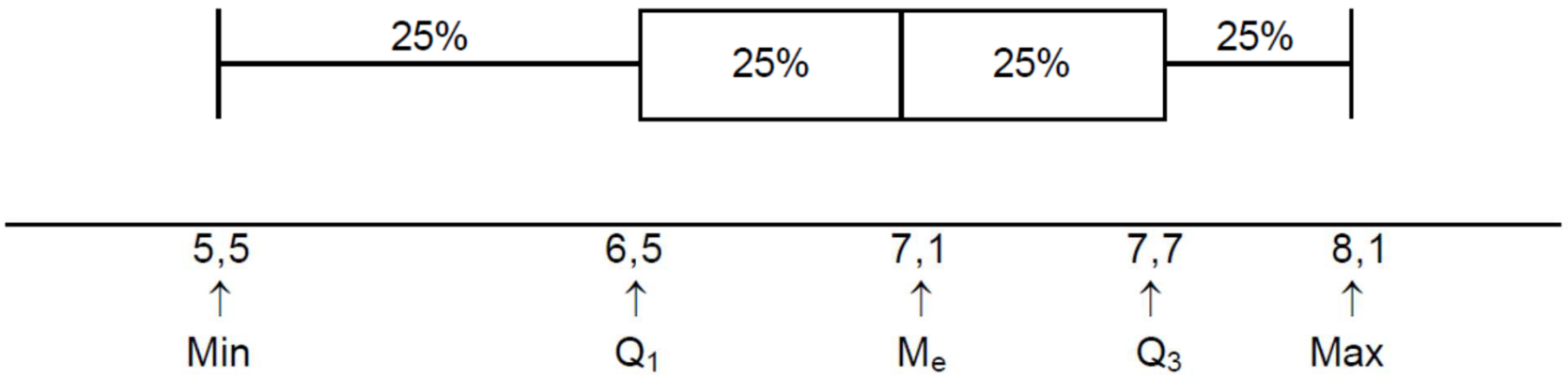
Il box plot di una distribuzione è un grafico caratterizzato da tre elementi principali:

- una linea che indica la posizione della **mediana** della distribuzione
- un rettangolo (box) i cui estremi sono determinati in base ai **quartili Q1 e Q3** della distribuzione e la cui altezza indica la variabilità dei valori prossimi alla mediana
- due segmenti che partono dal rettangolo i cui estremi sono determinati in base ai **valori minimo e massimo** della distribuzione



	y1	q1	m	q3	yn
voto esame di statistica	18	23	26	28	30

INTERPRETAZIONE DEL BOXPLOT



INTERPRETAZIONE DEL BOXPLOT

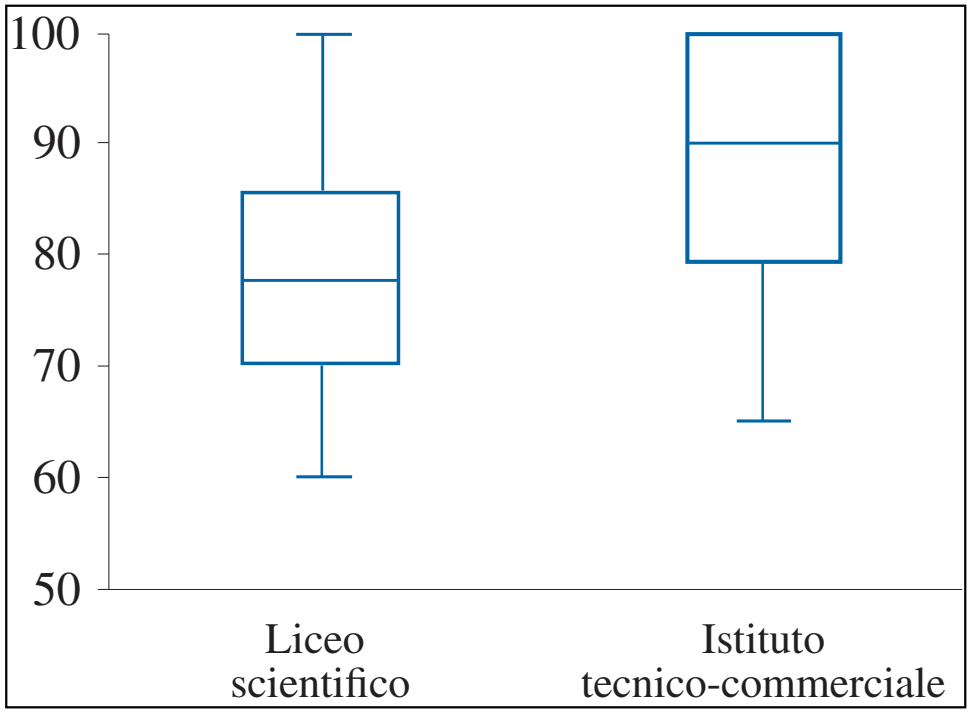
Il box-plot è utile perché riassume mediante pochi numeri molte informazioni su una distribuzione di frequenze.

- La mediana riassume la tendenza centrale della distribuzione.
- I quartili danno un'indicazione sulla variabilità, perché con essi si calcola lo scarto interquantile (misura più robusta del campo di variazione).
- La posizione della mediana rispetto ai quartili fornisce altre utili informazioni (in particolare sulla asimmetria della distribuzione).
- Gli estremi forniscono indicazioni non solo sul valore massimo e valore minimo ma soprattutto sull'eventuale presenza di dati con caratteristiche anomale

Diagramma a scatola: esempio

Aspetti salienti delle due distribuzioni

- **Intensità media** (*mediana*)
- La **misura della variabilità** (*campo di variazione* e *differenza interquartilica*)
- La presenza di **asimmetria** positiva o negativa che si desume dal raffronto delle distanze del primo e del terzo quartile dalla mediana



Voti riportati all'esame di Stato in un campione di studenti del corso di laurea in Economia aziendale distinti per scuola frequentata

Valori anomali ed estremi

Un dato è **anomalo** se:

- è maggiore del valore $Q3+1.5\Delta_q$
- è minore del valore $Q1-1.5\Delta_q$

Un dato è **estremo** (estremamente anomalo) se

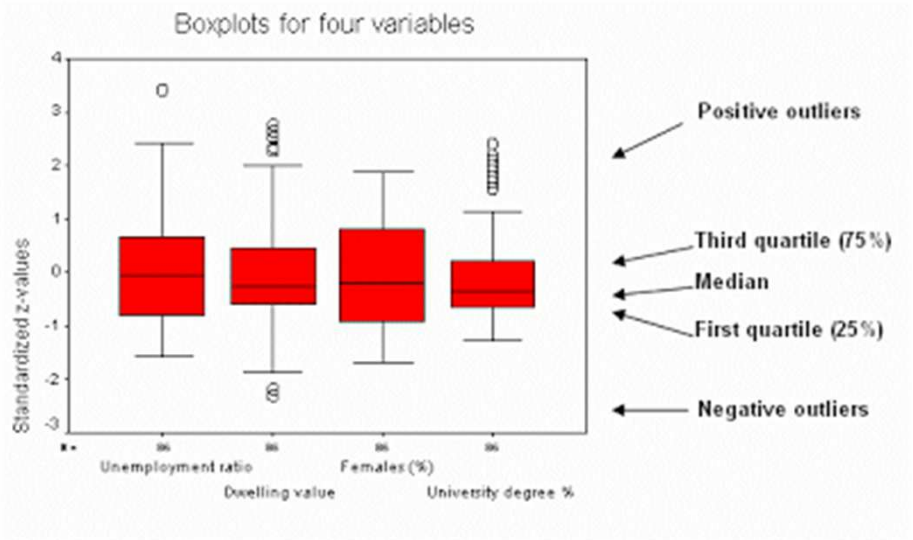
- è maggiore del valore $Q3+3\Delta_q$
- è minore del valore $Q1-3\Delta_q$

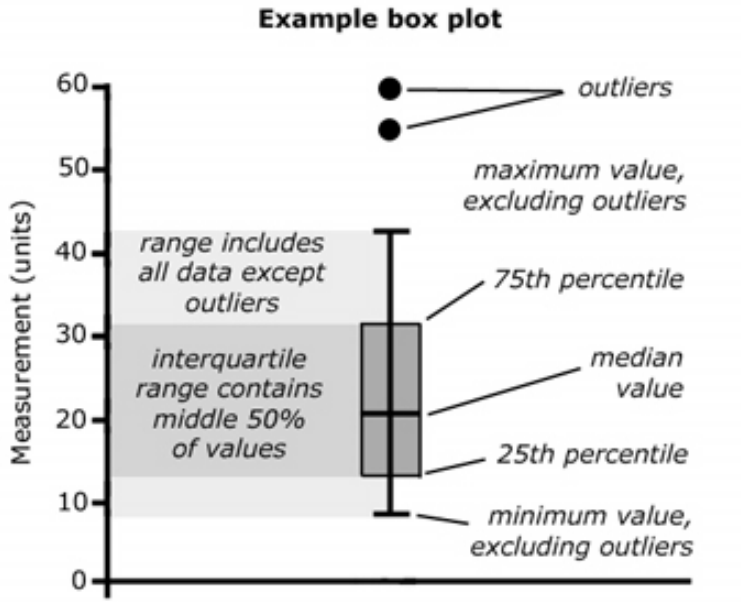
Boxplot e valori anomali ed estremi

I valori anomali ed estremi possono essere rappresentati nel boxplot.

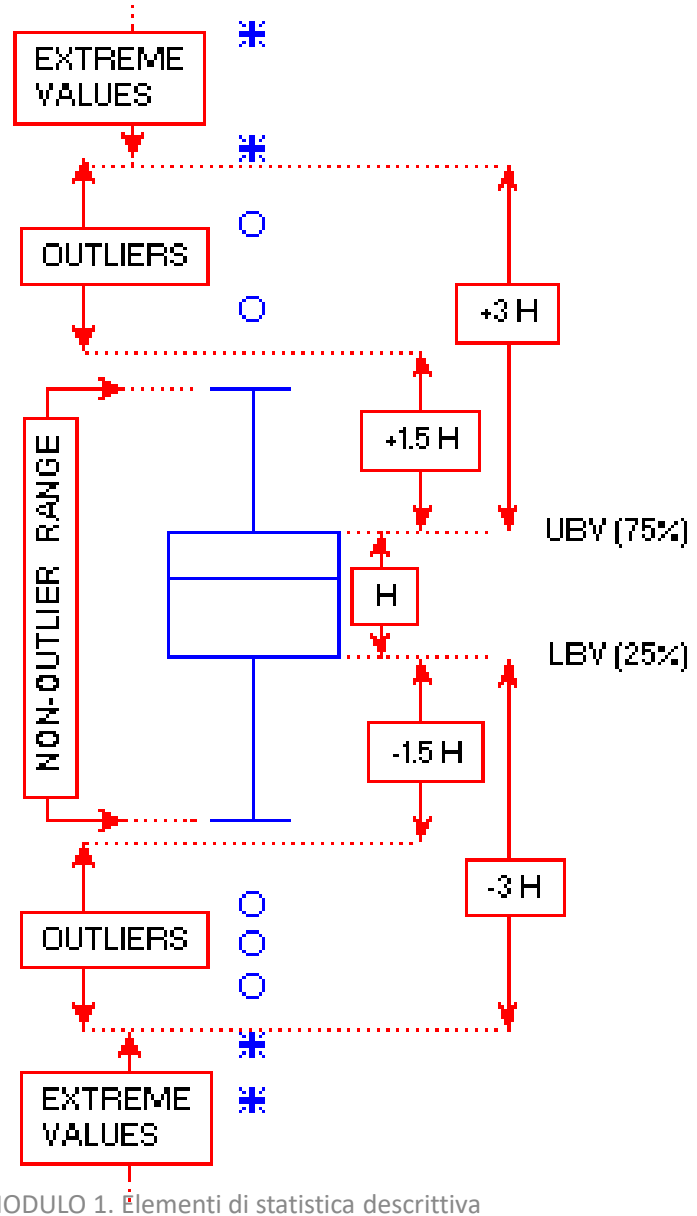
In questo caso i «baffi» (segmenti che partono dal rettangolo) hanno come estremi

- il valore più grande fra il minimo e $Q1-1.5\Delta_q$
- Il valore più piccolo fra il massimo e $Q3+1.5\Delta_q$

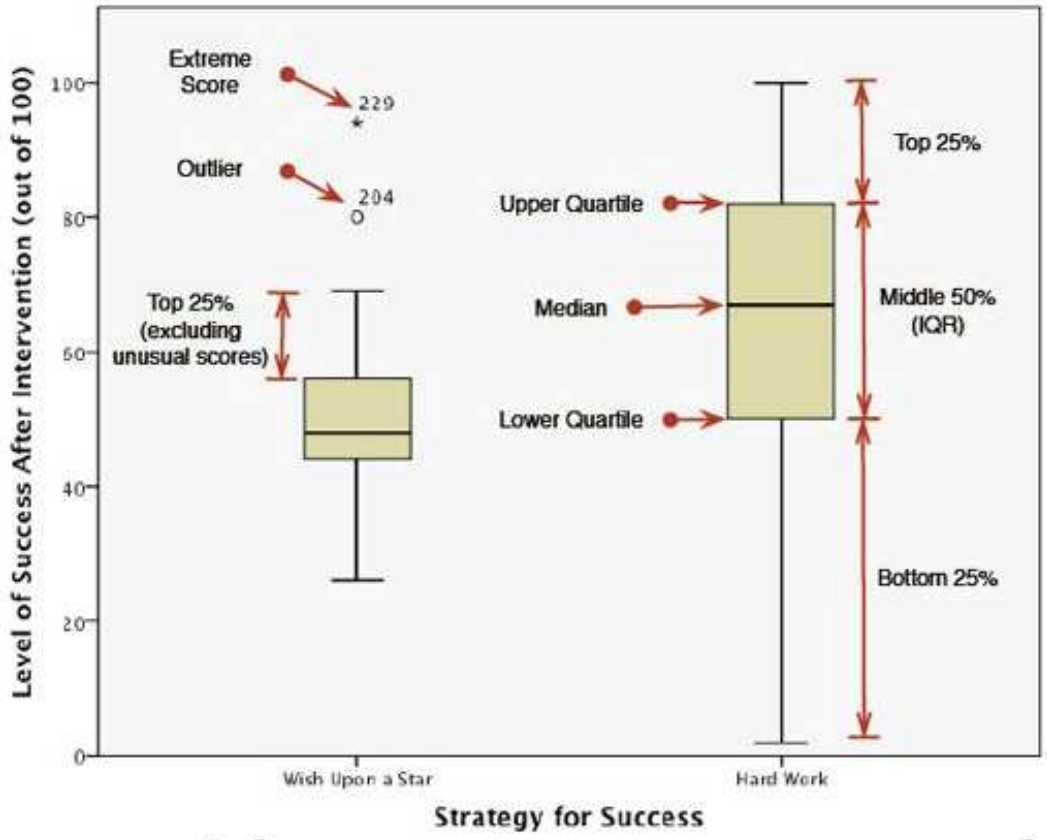




Boxplot e valori anomali ed estremi



Boxplot of success scores 5 years after implementing a strategy of working hard or wishing upon a star



Domande (1)

- Se hai la varianza come calcoli la deviazione standard?
  - 1. dividi per la numerosità del collettivo
  - 2. elevi la varianza al quadrato
  - 3. estrai la radice quadrata della varianza
  - 4. sottrai la media
  - 5. calcoli il reciproco della varianza

E se hai la devianza, quali delle operazioni elencate sopra effettui per calcolare la deviazione standard?

Se si vuole calcolare di quanto in media i valori sono distanti dalla media aritmetica, perché non si calcola semplicemente la media aritmetica delle differenze dei valori osservati rispetto alla media?

Domande (2)

Perché si ha bisogno di una misura della variabilità della distribuzione?

Cosa indica una deviazione standard uguale a zero?

I voti riportati da uno studente in 6 esami sono 18, 25, 30, 26, 27, 30. Un altro studente agli esami ha preso 26, 25, 24, 26, 26, 29.

Quale dei due studenti ha un rendimento maggiormente costante?

Confrontare le medie e le deviazioni standard

Trova il range ed il range interquartilico dei seguenti valori:

31, 16, 39, 15, 18, 41, 24, 26, 21, 33, 9, 12

Domande (3)

Disegna il boxplot del tasso di abbandono, evidenziando anche i valori anomali ed estremi

REGIONI	%
Piemonte	6,9
Valle d'Aosta/Vallée d'Aoste	10,8
Lombardia	6,6
Trentino-Alto Adige/Südtirol	3,2
Veneto	4,3
Friuli-Venezia Giulia	4,6
Liguria	7,5
Emilia-Romagna	6,6
Toscana	7,4
Umbria	4,7
Marche	4,8
Lazio	5,8
Abruzzo	6,1
Molise	5,7
Campania	9,3
Puglia	5,3
Basilicata	5,3
Calabria	5,3
Sicilia	9,2
Sardegna	10,4
Bolzano/Bozen	2,9
Trento	3,6

** Valori %, anno di riferimento 2012*

Domande (4)

	Drug trafficking recorded by the police, 2006–12							
	Number							
	2006	2007	2008	2009	2010	2011	2012	*totale*
Germany	64,865	64,093	55,905	50,965	49,622	50,791	47,667	***383,908***
France	5,792	5,797	6,128	6,007	5,869	5,928	4,821	***40,342***
Italy	32,306	34,439	34,082	34,101	32,761	34,034	33,852	***235,575***
Totale	**102,963**	**104,329**	**96,115**	**91,073**	**88,252**	**90,753**	**86,340**	***659,825***

- Confronta la variabilità delle denunce per traffico di droga nei tre Stati